

Claims

- [c1] A method for automatically classifying unclassified documents, comprising the steps of:
- a. processing, on a first processing system, a plurality of sample documents to identify a plurality of sample document feature sets of potentially duplicated and significant sample document features, whereby each sample feature set is associated with one of said plurality of sample documents;
 - b. electronically associating with each of said plurality of sample document feature sets a set of at least one manually selected document annotation values, whereby said document annotation values each represent a subjective classification of one of said plurality of sample documents with which said document annotation values are individually associated;
 - c. electronically associating with each of said plurality of sample document feature sets a set of at least one manually selected document feature annotation values, whereby said document feature annotation values each represent a subjective classification of one of a plurality of sample document features with which said document feature annotation values are individually associated;

d. processing, on a second processing system, an unclassified document to identify a set of potentially duplicated and significant unclassified document features;

e. comparing, on said second processing system, said set of potentially duplicated and significant unclassified document features to each of said sample document feature sets, inclusive of said document annotation values and said document feature annotation values associated with each of said sample document feature sets;

f. determining which of said plurality of sample document feature sets shares in common with any of the features comprising an unclassified document feature set a largest weighted quantity of features subjectively classified and annotated as significant, whereby a most significantly resembling sample document may be determined; and

g. outputting a significant similarity measurement value and a classification value for said unclassified document according to a weighted ratio of matching significant features of said most significantly resembling sample document as compared to all of said significant features of said most significantly resembling sample document.

[c2] The method of claim 1 wherein the documents to be classified are electronic messages such as email messages, wireless text messages, or instant messages.

- [c3] The method of claim 1 wherein said documents to be classified are electronic resume files.
- [c4] The method of claim 1 wherein said documents to be classified are HTML files or Web page files.
- [c5] The method of claim 1 wherein said documents to be classified are text files, regardless of the existence or lack of formatting information.
- [c6] A method for automatically classifying unclassified documents, comprising the steps of:
- a. registering, on a first processing system, each of said plurality of sample documents representative of at least one of a plurality of document classifications;
 - b. parsing each of said plurality of sample documents into at least one of a plurality of partial document content features according to a set of document parsing rules;
 - c. selectively decoding, removing and discarding from each of said sample documents, according to a set of document content decoding and removal rules, at least one of a plurality of said partial document content features, or portions of partial document content features, whereby any of said partial document content features that are considered insignificant for document classifica-

tion purposes or are considered to be obfuscating content that exists to subvert said document classification process may be removed;

d. determining and recording, by a manual document review and electronic annotation process, at least one of a plurality of subjective classifications of each of said plurality of sample documents, whereby at least one of a plurality of subjective classification labels are associated with each of said sample documents;

e. determining and recording, by a manual document review and electronic annotation process, at least one of a plurality of subjective classifications of each of said plurality of partial document content features of each of said sample documents, whereby at least one of a plurality of subjective classification labels are associated with each of said sample document's partial content features;

f. storing for each annotated sample document, on said first processing system, an annotated sample document record, inclusive of said sample document's content, said set of partial document content features, a set of unique digests of each partial content feature, at least one of said document annotation values, at least one of said plurality of said document feature annotation values, and other document attribute data;

g. storing, on said second processing system, a copy of

each of said annotated sample document records;

h. parsing, on a second processing system, an unclassified document into at least one of said plurality of partial document content features and selectively removing and discarding portions of said unclassified document's content in a manner consistent with steps 6b and 6c above;

i. querying said second processing system using said unclassified document's residual partial document content features or unique digests thereof and returning a list of all partially resembling sample documents which share in common at least one of a plurality of matching partial document content features with said unclassified document, subject to a requirement that any of said partial document content features that match are also subjectively classified and annotated as significant in any of said sample documents.

j. calculating a set of ratios of characters comprising said unclassified document's partial document content features that match said significant partial document content features contained in each of said partially resembling sample documents in said set of partially matching sample documents, as compared to a count of total characters comprising said significant partial document content features found in said partially resembling sample documents, resulting in a set of significant partial document content feature similarity scores;

k. comparing the highest of said scores to a predetermined document similarity threshold value; and
l. assigning said unclassified document said document similarity score and a classification value matching said subjective classification of said most closely resembling sample document if said document similarity score exceeds said predetermined threshold value, otherwise assigning said unclassified document a null or non-matching classification.

- [c7] The method of claim 6 wherein said plurality of partial document content features are comprised of non-overlapping character sequences or subsequences.
- [c8] The method of claim 6 wherein said plurality of partial document content features may be limited in length, including a minimum and maximum character length.
- [c9] The method of claim 6 wherein said plurality of partial document content features may be adjusted in length by truncation and concatenation with an adjacent partial document content feature of a same type.
- [c10] The method of claim 6 wherein index values may be associated with said plurality of partial document content features representing an order of appearance of said partial document content features in said document.

- [c11] The method of claim 6 wherein said partial document content features may be comprised of character sequences or subsequences separated by line break symbols, formatting tags and arbitrarily selected boundary types.
- [c12] The method of claim 6 wherein one of a plurality of partial document content feature types may be defined as any character sequence or subsequence conforming to a pattern of a hypertext link.
- [c13] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as any character sequence conforming to a pattern of a consistently recognizable portion of a hypertext link.
- [c14] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as an attached file's contents.
- [c15] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as a linked file's contents.
- [c16] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as an attached file's metadata.

- [c17] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as a linked file's metadata.
- [c18] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as a call-to-action character sequence or subsequence.
- [c19] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as an insignificant character sequence or subsequence.
- [c20] The method of claim 6 wherein one of said plurality of partial document content feature types may be defined as an executable program code character sequence or subsequence.
- [c21] The method of claim 6 wherein more than one method of partitioning said document into partial document content features may be used to produce more than one set of partial document content features, whereby more than one method of measuring document similarity may be employed.
- [c22] The method of claim 6 wherein decoding of any encoded partial document content features uses a distinct set of decoding rules for said partial document content fea-

tures of specified types and of specified document feature encoding types.

- [c23] The method of claim 6 wherein decoding and removal of potentially insignificant or obfuscating content from any partial document content features uses a distinct set of content removal rules for said partial document content features of specified types.
- [c24] The method of claim 6 wherein said calculation of said similarity score ratio employs weights for each of said partial document content features that are proportional to the number of text characters comprising each of said partial document content features.
- [c25] The method of claim 6 wherein the numbers of characters used to assign weights for partial document content features exclude characters which have been removed.
- [c26] The method of claim 6 wherein a plurality of similarity threshold values may be applied to determine document similarity, whereby a specific similarity threshold value may be applied conditionally, depending upon an attribute of said document, such as said document's total character length.
- [c27] The method of claim 6 wherein a first unclassified document having fewer than a predetermined number of

characters is evaluated against a higher similarity score threshold value than a second unclassified document having a number of characters greater than a predetermined number of characters.

- [c28] A method for automatically identifying in a document a set of potentially duplicated and significant document features, comprising the steps of:
- a. parsing said document into at least one of a plurality of said partial document content features according to a set of document parsing rules;
 - b. selectively removing and discarding from said sample document, according to a set of document content removal rules, at least one of a plurality of said partial document content features, or portions of said partial document content features, that are considered insignificant for document classification purposes or are considered to be obfuscating content that exists to subvert said document classification process, whereby any remaining content may be considered potentially duplicated and significant.
- [c29] The method of claim 28 comprising the step of removing partial document content features whereby content of different partial document content features types are removed according to different rules and at different stages in a sequence of content removal steps, whereby

content removal rules may be invoked conditionally depending upon said stage of processing and said partial document content feature type to be processed.

- [c30] A method of excluding from consideration in a document similarity measurement process semantically insignificant or obfuscating partial document content features contained within sample documents, comprising the steps of:
- a. selecting and recording, by a manual document review and electronic annotation process, at least one of a plurality of subjective classification values of each of said plurality of partial document content features of said sample documents, wherein at least one of said plurality of subjective classification values are bound to a record of each of said sample documents' partial content features;
 - b. assigning a numerical weight of zero to any of said partial document content features which are labeled with a classification value indicating that said partial document content features are of a semantically insignificant or obfuscating content classification; and
 - c. including said zero-weighted classification values in said similarity measurement process steps that apply said weights to be assigned to each of said partial document content features comprising said sample docu-

ments.

- [c31] A method of preventing the submission of a new sample document to a manual document review and annotation processing system when said new sample document is an exact or significantly partial duplicate of a previously submitted, reviewed and retained sample document, comprising the steps of:
- a. parsing said new sample document into at least one of said plurality of partial document content features according to said set of document parsing rules;
 - b. selectively removing and discarding from said new sample document, according to said set of document content removal rules, at least one of a plurality of said partial document content features, or portions of said partial document content features, that are considered insignificant for document classification purposes or are considered to be obfuscating content that exists to subvert said document classification process;
 - c. querying said first processing system using said new sample document's residual partial document content features or unique digests thereof and returning a list of all partially resembling existing sample documents which share in common at least one of a plurality of matching partial document content features with said new sample document, subject to said requirement that any of said

partial document content features that match are also subjectively classified and annotated as significant in any said existing sample documents;

d. calculating a set of said ratios of characters comprising said new sample document's partial document content features that match said significant partial document content features contained in each of said partially resembling existing sample documents, as compared to said total characters comprising said significant partial document content features found in said partially resembling sample documents, resulting in a set of significant partial document content feature similarity scores;

e. comparing the highest of said scores to said predetermined document similarity threshold value; and

f. accepting submission of said new sample document if said similarity score falls below a predetermined similarity score threshold value; and

g. discarding said new sample document if said similarity score equals or exceeds said predetermined similarity score threshold value, whereby said new sample document is excluded from said manual document review process due to its significant measured similarity to one of said plurality of existing sample documents.

[c32] A method of calculating a measure of similarity between two sets of partial document content features that ad-

justs for differences in relative length of partial document content features, comprising the steps of:

a. determining which of said set of partial document content features of a first document match any of said set of partial document content features of a second document, wherein said partial document content features are extracted from each of said documents according to the same method;

b. calculating a similarity score, wherein a similarity score is a ratio of said number of characters contained in matching partial document content features divided by said total number of characters in all of said partial document content features comprising said first document.

[c33] The method of claim 32 comprising the step of detecting and deleting any of said partial document content features that match one of a plurality of common partial document content features.

[c34] The method of claim 32 comprising the step of removing partial document content features, or portions thereof, according to a set of content removal rules that are dependent on said type of partial document content feature, before counting characters contained in said partial document content feature.

[c35] A method of automatically determining the topical clas-

sification of a document, comprising the steps of:

- a. determining that at least a minimum quantity of partial document content features of an unclassified document match any of a set of said partial document content features of a previously classified document;
- b. determining that at least a minimum weighted relative quantity of said matching partial document content features of said previously classified document are individually classified as being indicative of said previously classified document's topical classification;
- c. assigning a topical classification of said previously classified document to said unclassified document.

[c36] The method of claim 35 wherein the method of weighting said quantity of partial document content features is based on a count of characters comprising each of said partial document content features.

[c37] The method of claim 35 wherein said count of characters comprising each of said partial document content features is calculated after completing a partial document content removal process to eliminate insignificant or obfuscating content.

[c38] A method of selecting and collecting unclassified documents distributed in a network that may serve as samples of similar documents to be classified, comprising

the steps of:

- a. storing, for each unclassified or non-specifically classified document distributed in a network, profiles comprised of each document's partial document content features;
- b. deriving, for a first new document distributed within a network, a profile comprised of said first new document's partial document content features;
- c. calculating a measure of similarity of said first new document's profile relative to each of said existing unclassified or non-specifically classified document profiles;
- d. classifying as partially duplicated said first new document for which at least a predetermined minimum measure of similarity is calculated with respect to its profile as compared to any of said existing unclassified or non-specifically classified document profiles;
- e. retaining as a candidate new sample document said first partially duplicated document copy and its profile.

[c39] The method of claim 1 wherein said sample documents are collected and processed by a service provider.

[c40] The method of claim 1 wherein said sample documents are collected and processed by an administrator of a user network.

[c41] The method of claim 1 wherein said manual review of said sample document results in recording a subjective classification of any of said partial document content features that are insignificant for document similarity detection purposes.

[c42] The method of claim 1 wherein said manual review of said sample document results in recording a subjective classification of any of said partial document content features that are indicative of said sample document's topic classification.